

# SUR L'APPORT DES *BIG DATA* : UN EXEMPLE TIRÉ DE LA SOCIOLOGIE DU COUPLE

Marie Bergström<sup>1</sup>

<sup>1</sup> Institut national d'études démographiques (Ined), 133 boulevard Davout, 75020 PARIS,  
marie.bergstrom@ined.fr

## Résumé

On observe actuellement un débat sur l'intérêt et les limites des « *big data* » en sciences sociales. La communication se veut une contribution à cette discussion. En partant d'un exemple empirique précis, il s'agit d'illustrer quelques-unes des opportunités de recherche offertes par ce nouveau type de données. L'exemple proposé s'inscrit dans la sociologie du couple. Il porte sur une étude de l'écart d'âge entre conjoints, mobilisant des données issues d'un site de rencontres sur Internet. L'analyse porte, d'une part, sur les préférences d'âge affichées par les utilisateurs dans leur profil et, d'autre part, sur leurs comportements de contacts (qui échange avec qui ?). On souligne l'apport de cette méthodologie par rapport à l'usage habituel des enquêtes par questionnaire relatives à des couples déjà constitués. Contrairement à cette dernière approche – qui étudie la formation des couples *ex post facto* – l'analyse des sites de rencontres permet d'observer le processus d'appariement des partenaires. Ce changement de perspective donne des résultats nouveaux. Alors que les données d'enquête indiquent que l'écart d'âge est surtout désiré par les femmes, les analyses du site de rencontres montrent qu'il est aussi recherché par les hommes, surtout lors d'une remise en couple. Plus généralement, l'étude indique que les couples résultent d'un arbitrage entre les préférences masculines et féminines qui divergent plus souvent qu'elles ne concordent. À partir de cet exemple concret, la présentation souligne donc les perspectives de recherches nouvelles offertes par ces « *big data* » et invite à continuer le débat à leur sujet.

**Mots-clés.** Big data, enquête par questionnaires, couple, écart d'âge, sites de rencontres

## Une illustration empirique de l'apport des données « organiques »

Depuis une dizaine d'années, on observe un intérêt croissant en sciences sociales pour les *big data*, ou plus précisément pour les données « organiques » (*organic data* en anglais). Ce terme, forgé par Robert M. Groves (ancien directeur d'*US Census*), désigne des données qui sont produites dans d'autres objectifs que la recherche mais qui sont potentiellement exploitables par celle-ci (Groves, 2011). Il s'agit notamment de registres administratifs, de données d'entreprises ou de traces d'usages d'Internet. Sans être forcément « *big* », ces matériaux empiriques se distinguent des données d'enquête qui sont, elles, produites dans l'objectif explicite de l'analyse scientifique (*designed data* en anglais).

L'intérêt pour ces données trouve, en sciences sociales, au moins deux origines. D'une part, on espère y trouver une solution à la baisse des taux de réponse aux enquêtes par questionnaire. Cette tendance, qui affaiblit la représentativité des enquêtes et en augmente les coûts, a conduit les chercheurs à se tourner vers d'autres sources d'information. D'autre part, on y voit des nouvelles perspectives d'analyse. Bien souvent, les données dites « organiques » enregistrent des *traces de pratiques* (recours aux soins, usage de transports publics, usages d'Internet...) plutôt que des *déclarations de pratiques* comme c'est le cas des questionnaires. Plus qu'une source d'information alternative, ces données sont alors utilisées pour produire des connaissances nouvelles sur le monde social.

C'est ce deuxième usage des données « organiques » qui est le point de départ de cette communication. On propose de mobiliser des données issues d'Internet pour revisiter un objet classique de la sociologie, à savoir la formation des couples. Plus précisément, la communication présente une étude sur l'écart d'âge entre conjoints à partir de données extraites d'un site de rencontres en ligne. L'objectif est double. La volonté est tout d'abord de contribuer au débat actuel sur l'intérêt et les limites des *big data* en sciences sociales : en partant d'un exemple empirique précis, on espère illustrer quelques-unes des opportunités de recherche offertes par ces nouvelles données. Ce faisant, il s'agit aussi d'apporter des connaissances nouvelles sur cette tendance forte qu'est l'écart d'âge entre partenaires.

On revient ici rapidement sur le contexte et les objectifs de la présentation, les données mobilisées et enfin les principaux résultats de cette étude en cours.

## **Sites de rencontres : un nouveau site d'observation des comportements conjugaux**

*Contexte : traditions & innovations dans l'étude de la formation des couples*

La recherche sur la formation des couples repose très largement sur des enquêtes par questionnaire dont les répondants sont en union libre ou mariés. Ces enquêtes sont mobilisées non seulement pour décrire la nature des couples formés (degré et évolution de l'homogamie par exemple) mais aussi pour étudier les « goûts » conjugaux et les logiques d'appariement des partenaires (comme l'homophilie). Dans ce deuxième cas, l'approche est nécessairement indirecte et rétrospective : la formation des couples n'est pas étudiée en tant que telle mais analysée *ex post facto*. Autrement dit, les pratiques et les préférences ne sont pas observées mais inférées à partir d'unions déjà constituées. Il en résulte que notre compréhension du « marché matrimonial » repose davantage sur des modèles d'acteurs que sur des études empiriques. Comme le remarque Matthijs Kalmijn (1998), la sociologie du couple souffre à ce titre d'un certain vide entre théorisation et travail empirique.

Les sites de rencontres peuvent permettre de pallier ce vide. D'abord, ces services enregistrent les préférences amoureuses des utilisateurs. Ensuite, ils captent les comportements de contacts (qui discute avec qui ?) et permettent ainsi d'observer le processus d'appariement des usagers. Lorsqu'extraites, codées et analysées par le chercheur, ces informations peuvent apporter des indications précieuses sur les comportements conjugaux (Hitsch, Hortacsu, & Ariely, 2010 ; Skopek, Schulz, & Blossfeld, 2011 ; Potârca & Mills, 2015). Contrairement aux données d'enquête, elles placent le curseur *au moment* de la sélection amoureuse et captent aussi bien l'*élimination* que la *sélection* des partenaires. Cette nouvelle approche – méthodologique et empirique – peut alors apporter de nouvelles connaissances.

C'est ce qu'on tentera de montrer en s'intéressant plus spécifiquement à l'écart d'âge entre conjoints hétérosexuels. Cet objet classique de la sociologie du couple est essentiellement abordé à partir de données d'enquêtes. Celles-ci informent sur l'ampleur et l'évolution de la différence d'âge dans les unions mais peinent à rendre compte de la manière dont cet écart *se produit*. Est-ce que ce sont les femmes, les hommes ou les deux sexes qui désirent cette asymétrie ? Les réponses à cette question diffèrent selon les sources utilisées. On montrera que l'écart d'âge entre partenaires trouve des explications différentes selon qu'on mobilise des données d'Internet ou une enquête par questionnaire.

## *Méthodologie : mise en perspective des données d'enquêtes & des données « organiques »*

La présentation propose donc une mise en perspective des données d'enquête et des données d'Internet au sujet de la différence d'âge entre partenaires hétérosexuels. À cet objectif, on mobilise tout d'abord l'enquête *Étude des parcours individuels et conjugaux* (Épic), conduite à Ined en 2013-2014 sous la direction de Wilfried Rault et Arnaud Régnier-Loilier. Consacrée à la vie affective des Français, l'enquête interroge les répondants sur chacune des « relations amoureuses importantes », en cours ou vécues par le passée. 7 825 personnes âgées de 26 à 65 ans ont répondu à l'enquête et, au total, 14 699 relations ont été décrites.

Cette étude permet d'abord de mesurer *l'écart d'âge moyen* entre conjoints et son variation selon les caractéristiques des couples. Mais elle donne aussi des indications sur les *préférences d'âge* des femmes et des hommes. Partant du principe que le couple ne reflète pas forcément les préférences des (deux) partenaires, on a invité les répondants à se positionner par rapport à un scénario conjugal hypothétique. Il leur a été demandé s'ils auraient « facilement accepté l'idée » d'être avec quelqu'un plus jeune ou plus âgé qu'eux (de 5 ans). Cette question vise à cerner les attitudes des deux sexes vis-à-vis de l'écart d'âge traditionnel et on la mobilise dans cet objectif.

Les données « organiques » mobilisées dans le cadre de cette étude proviennent d'un des plus grands sites de rencontres en France en termes de nombre d'inscrits. Dans le cadre d'un accord négocié avec l'entreprise propriétaire du site, j'ai obtenu en 2014 des informations de deux types.

Une première base de données comporte l'intégralité des profils d'utilisateurs enregistrés sur le site depuis son lancement en 2002 (N : 16M). Elle contient l'ensemble des informations affichées dans ces profils (à l'exception des pseudonymes et des photographies). Pour cette présentation, on tiendra surtout compte des caractéristiques sociodémographiques renseignées (sexe, âge, niveau de diplôme et parentalité) et des préférences d'âge déclarées. Sur le site en question, les utilisateurs sont obligés de renseigner une tranche d'âge souhaitée pour les partenaires potentiels allant de 18 à 99 ans. L'étude présentée repose en partie sur une analyse statistique de ces préférences.

Une deuxième base de données comporte des métadonnées quant à l'ensemble des emails envoyés sur le site depuis 2011 (N : 206M). Elle ne renseigne pas sur le contenu des écrits mais seulement sur le contexte de l'échange : identifiants de l'expéditeur et du destinataire ainsi que la date et l'heure de l'envoi. Recoupée avec les données de profil, cette information permet d'analyser les logiques d'appariement des utilisateurs. On centrera ici l'analyse sur les comportements de contacts par rapport à l'âge des interlocuteurs.

## *Résultats : deux approches méthodologiques pour des résultats différents*

L'enquête Épic confirme les grandes tendances quant à l'écart d'âge entre conjoints, déjà observées dans d'autres enquêtes. En moyenne, les hommes ont environ deux ans de plus que leur conjointe<sup>1</sup>. Cet écart se creuse au fil des unions pour les hommes alors qu'il se réduit avec le rang d'union pour les femmes. Surtout, Épic confirme des résultats antérieurs quant aux attitudes des deux sexes vis-à-vis de l'écart d'âge en faveur de l'homme. Tout comme l'enquête sur la *Formation des couples*, conduite en 1983-1984, elle indique une plus forte préférence pour cet écart chez les femmes que chez les hommes (Bozon, 1990). En 2013-2014, les hommes sont toujours plus nombreux que les pairs féminins à déclarer qu'ils accepteraient « facilement » un couple où la femme serait l'ainée. L'asymétrie des âges dans les couples hétérosexuels apparaît donc comme le produit des préférences des femmes en premier lieu.

L'analyse des données du site de rencontre donne, elle, des résultats différents par

---

<sup>1</sup> Pour les générations les plus récentes, nées entre 1978 -1987, l'homme avait en moyenne 2,3 ans plus que sa conjointe (Rault & Régnier-Loilier, 2015).

rapport aux résultats d'enquête. Tout d'abord, elle indique que les attitudes vis-à-vis de l'écart d'âge varient au cours de la vie. En début de l'âge adulte, ce sont bien les utilisatrices qui expriment la préférence la plus forte pour un écart d'âge traditionnel. Cela est visible aussi bien dans les critères d'âges affichés dans les profils que dans les comportements de contacts : une large majorité des mails envoyés par les utilisatrices le sont à des utilisateurs plus âgés. Cependant, chez les usagers âgés de 40 ans et plus ce sont au contraire les utilisateurs qui manifestent le désir le plus fort pour un écart en leur faveur. Les critères d'âge *déclarés*, et plus encore les préférences *manifestées* (par le choix des interlocutrices) indiquent qu'aux âges qui correspondent aux deuxièmes unions, les hommes tiennent plus que les femmes à cette asymétrie traditionnelle. Les résultats quant au scénario hypothétique (proposé dans l'enquête par questionnaire) tranchent donc avec les pratiques observées *in situ* sur les sites de rencontres.

Aussi, les données du site indiquent que la production de l'écart d'âge est un processus complexe. Alors que les couples sont souvent considérés comme des « préférences révélées » (*cf.* Samuelson, 1938) – les caractéristiques du partenaire féminin refléteraient les désirs masculins et *vice versa* – les « paires » formées en ligne indiquent que les couples résultent plutôt d'un *arbitrage* entre les désirs des deux sexes. Sur le site de rencontres étudié, les préférences d'âge des utilisatrices et des utilisateurs divergent plus souvent qu'elles ne concordent. L'écart d'âge entre les interlocuteurs traduit un *point de rencontre* entre des désirs qui contrastent par ailleurs. Autrement dit, le couple apparaît comme le produit de deux aspirations accommodées. Ce jeu de négociations et de concessions à l'origine de la formation des couples ne peut être observé *ex post facto*. Seule une approche inductive permet d'en observer les mécanismes.

\*  
\* \*

Depuis le milieu du  $xx^e$  siècle, les enquêtes par questionnaire sont au cœur de la recherche quantitative en sciences sociales. Elles sont des puissants outils d'objectivation mais comportent aussi des limites. D'une part, elles reposent sur un accès indirect aux informations recherchées : les enquêtes se basent sur des déclarations et non des observations, laissant ainsi à la charge de l'enquêté de rendre compte de ses pratiques et de ses attitudes. D'autre part, les enquêtes par questionnaire enregistrent aisément des états (comme les couples) mais peinent cependant à capter des processus (comme l'appariement des partenaires). Face à ces limites, les données dites « organiques » s'avèrent très complémentaires. C'est particulièrement le cas des données d'Internet qui recueillent des traces d'action et enregistrent souvent des flux. Bien qu'elles soient non représentatives et généralement pauvres en information, ces informations peuvent éclairer certaines zones d'ombres laissées par les traditionnelles enquêtes par questionnaire.

L'étude d'un site de rencontres se veut une illustration de l'utilité de cet usage précis et ponctuel des données d'Internet afin de pallier un point aveugle des études habituelles. Alors que les enquêtes sur la formation des couples peinent à rendre compte des aspirations amoureuses *en amont* de la rencontre et le processus d'appariement des partenaires, les sites permettent d'observer précisément cela. Ils offrent ainsi un terrain pour tester des hypothèses et peuvent *in fine* participer à la consolidation empirique des théories du choix amoureux.

## Bibliographie

- [1] Groves, R. M. (2011), Three eras of survey research, *Public Opinion Quarterly*, 75(5), 861-871.
- [2] Kalmijn, M. (1998), Intermarriage and Homogamy: Causes, Patterns, Trends, *Annual Review of Sociology*, 24, 395-421.
- [3] Hitsch, G., Hortacsu, A., et Ariely, D. (2010), Matching and Sorting in Online Dating, *The American Economic Review*, 100(1), 130-163.
- [4] Skopek, J., Schulz, F., et Blossfeld, H.-P. (2011), Who Contacts Whom? Educational

- Homophily in Online Mate Selection, *European Sociological Review*, 27(2), 180-195.
- [5] Potârcă, G., et Mills, M. (2015), Racial Preferences in Online Dating across European Countries, *European Sociological Review*, 31(3), 326-341.
- [6] Rault W., et Régnier-Loilier A. (2015), La première vie en couple : évolutions récentes, *Population & Sociétés*, 521, 1-4.
- [7] Bozon M. (1990), Les femmes et l'écart d'âge entre conjoints : une domination consentie. I. Types d'union et attentes en matière d'écart d'âge, *Population*, 45(2), 327-360.
- [8] Samuelson P. (1938), A Note on the Pure Theory of Consumers' Behaviour, *Economica*, 5(17), 61-71 .